Jestr Journal of Engineering Science and Technology Review 7 (4) (2014) 120- 124

JOURNAL OF
Engineering Science
and Technology Review

www.jestr.org

Research Article

A Fair QoE-Aware MIMO Scheduling Scheme Based on Dynamic Latency Boundary for Non-Real-Time Service

Lei Chen[1], Ping Wang[1,*], Fuqiang Liu[1] and Nguyen Ngoc Van[2]

[1]*Broadband Wireless Communications and Multimedia Laboratory, Tongji University, Shanghai 201804, China*
[2]*School of Electronics and Telecommunications, Hanoi University of Science and Technology, Hanoi 10999, Vietnam*

Received 25 May 2014; Accepted 23 September 2014

Abstract

In Multiple-input Multiple-output (MIMO) system, the scheduling schemes for guaranteeing quality of service (QoS) set fixed delay boundaries for data packets of special service. However, the utility function of the quality of user experience (QoE) is usually a continuous function mapping service latency to Mean Opinion Score (MOS). Based on a QoS provisioning scheme, we propose a QoE-aware scheduling scheme with dynamic latency boundary. We also propose a weight function, which is a logarithmic function of the mean opinion score (MOS) level, to guarantee fairness. Numerical results indicate that the proposed scheme achieves higher performance from the view of user experience and fairness. This scheme can be deployed in commercial networks because of its low computational complexity.

Keywords: Multiple-input/multiple-output (MIMO), quality-of-service (QoS), quality-of-experience (QoE), non-real-time service

1. Introduction

The non-real-time data services, such as web, social network service, http progress video, and instant messaging, have become more and more popular in mobile networks. This trend motivated telecom operators and equipment providers to design new scheduling strategies for guaranteeing the quality of user experience (QoE) and saving resources [1], [2], [3]. They deployed Content Distribution Networks (CDN) in mobile networks to reduce the latency in the core network. To decrease the capacity gap between the air interface and the core network, the Multiple-Input Multiple-Output (MIMO) technique has been applied to improve stability and the rate of wireless transmission. J.-B. Landre assessed this technique in realistic environments [4], [5], and S. Parkvall and A. Farajidana performed their test in a real, operational network [6], [7]. These tests demonstrate their commercial worth; however, MIMO also increases the complexity of the coding and scheduling.

Considerable studies have been focused on throughput capacity and quality-of-service (QoS) of MIMO systems. G. Caire, P. Viswanath, S. Vishwanath and H. Weingarten have characterized the capacity region of the Gaussian MIMO broadcast channel and proposed nonlinear scheduling schemes to maximize throughput, respectively [8]-[11]. T. Yoo proposed the "zero forcing beamforming" (ZFBF) scheme with lower time complexity [12]. To meet the packet delay requirements of different services, some QoS scheduling schemes have been proposed. D. Wu proposed the concept of "effective capacity" for mapping the throughput capacity to the QoS capacity [13]. J. Tang explored the trade-off between throughput and QoS provisioning through "effective capacity" and applied it into the MIMO system [14], [15]. Moreover, based on game theory, L. Zhong gave the optimal solution of throughput under delay and power constraints [16].

However, these schemes meet some challenges in commercial networks with non-real-time data services. First, because the function bridging the latency and the QoE is continuous, we cannot set a fixed delay boundary for data service packets. Next, many mobile network operators put all of the small data flow services into a common channel to save wireless resources. As the number of users increases, the base station (BS) cannot afford large time consumption for scheduling. Finally, a non-real time service might require a large number of time slots; the scheme needs to adapt to the variations in the channels and services.

In this paper, we propose a QoE-aware scheme based on a QoS scheme. We map the different QoE levels to the latency boundaries and adjust the expected QoE level according to the state of the service and the channel. The numerical results show that the proposed scheme results in higher average performance and fairness from the view of QoE with low computational complexity.

The rest of the paper is organized as follows: Section 2 describes the system model and the objective function from the view of user experience. Section 3 introduces the traditional schemes and the proposed schemes. Section 4 presents simulations and compares the effectiveness of the four schemes. The paper's conclusions are presented in section 5.

* E-mail address: 2010_dx_cl@tongji.edu.cn

2. System Model

We consider a MIMO system with N_t transmission antennas serving multiple users. The base station (BS) puts all of the non real time data services into one common channel, and each user has only one data service and uses one antenna. The type of service is web or short video. We assume that "Deep Packet Inspection" (DPI) can obtain perfect state information of the services, such as the service type, total size of the data flow.

2.1 MIMO System Model

We assume that the length of the time slot is smaller than that of the channel coherent time. Therefore, the signal received by user *k* can be given by

$$y_k = \underbrace{h_k v_k \sqrt{p_k} s_k}_{desired} + \underbrace{\sum_{j \neq k} h_k v_j \sqrt{p_j} s_j}_{\text{int}erference} + n_k \quad (1)$$

Where s_k is the data, $h_k \in C^{1 \times N_t}$ is the channel gain, $v_k = R^{N_t \times 1}$ is the steering vector, p_k is the allocated power, n_k is the Gaussian noise with variance $\mathrm{E}\left[n_k n_k^H\right] = \sigma^2$, and the term marked with interference is the in-cell interference. The Signal-Interference-plus-Noise-Ratio (SINR) can be given by

$$SINR_k = \frac{p_k \left| h_k v_k \right|^2}{\sum_{j \neq k} p_j \left| h_k v_j \right|^2 + \sigma^2} \quad (2)$$

with perfect CSIT, the rate of user k can be given by

$$r_{k_c} = \log_2 \left(1 + SINR_k\right) \times band \quad (3)$$

2.2 QoE of non-real-time Data Service

The "Mean Opinion Score" (MOS) is used to obtain the human user's view of the quality of service. This concept originated from voice tests [17]. The MOS of the data service can be obtained using some objective measurement methods formulated by the latency function [18]. The problem of maximizing the sum of the MOS of the users can be formulated as

$$\max_{r_k} \sum_{k=1}^{K} U_k \left(D_k\right) \quad (4)$$

Where U_k is the utility function of user *k* to obtain a MOS depending on the service type and D_k is the latency. For a current time slot, D_k can be given by

$$D_k = L \times \left(Past_k + F_k\right) \quad (5)$$

Where *L* is the duration of each time slot, $Past_k$ is the elapsed time from the beginning of the application, F_k is the estimated buffering time for the remaining data.

2.3 Traffic Model

Many scheduling schemes are performed using real-time traffic models or full buffer models characterized by a constant number of users and unlimited amounts of data flow for each user [19]. We adopt a full buffer model but with limited data in each data flow. We assume that the length of the data flow of each service is finite and services will stop after the reception is completed. Users might start new services after one service finishes. Moreover, the rate of the core network is faster than that of the air interface transmission. Therefore, the BS does not have to wait for the data from the remote server.

3. MIMO Scheduling Scheme for QoE

In this section we will give a brief review of "dirty paper code" (DPC) for MIMO systems and QoS schemes and describe the proposed QoE-aware scheme based on existing QoS schemes.

3.1 Review of ZF-DPC

The scheduling scheme called ranked known interference (RKI) is proposed for MIMO systems in reference [8]; this scheme is also known as zero force dirty paper code (ZF-DPC). ZF-DPC is an iterative scheme to obtain suboptimal throughput. Each iteration, the candidate users are appended to the sequence of selected users and the scheme applies Gram-Schmidt orthogonalization to the channel matrix. The candidate user with the highest SINR is selected. The time complexity of the scheme will increase sharply as the number of users increases. Moreover, the fairness of this scheme is very low.

This scheme needs $\left(N_t^2 \times K - N_t \times K\right)/2 - \left(N_t - 1\right) N_t (2N_t - 1)/6$ projections for user selection. Therefore, the time complexity of the algorithm is $\mathrm{O}\left(N_t^2 K\right)$, if *N* is treated as a constant, the time complexity is $\mathrm{O}(K)$, where *K* is the number of users and N_t is number of transmit antennas.

3.2 Review of the QoS Scheme Based on Game Theory

G. Song proposed a resource allocation method based on a gradient of the utility function with respect to the rate [19]. This method was used to explore the trade-off between QoS requirements and system throughput by L. Xingmin [20]. Based on the research results of [20], L. Zhong proposed a game theoretic QoS model in MIMO system for real time service [16]. The delay boundaries are modeled into weights using the "Nash Bargaining Solution" (NBS). Under the premise of meeting the latency constraint, the QoS-guaranteeing algorithm minimizes the delay for each user. The objective function is

$$\max_{r_k} \prod_{k=1}^{K} \left(D_k - q_k / r_k\right) \quad (6)$$

Where r_k is the possible rate combination, D_k is the latency boundary, and q_k is the length of the buffering queue data. This expression is equivalent to

$$\max_{r_k} \sum_{k=1}^{K} \ln\left(D_k - q_k / r_r\right) \quad (7)$$

Defining $U_r = \sum_{k=1}^{K} \ln\left(D_k - q_k / r_k\right)$, the solution can be found from the maximum value of

$$\nabla U_{\bar{r}}\left(t-1\right) \cdot \bar{r}\left(t\right) \quad (8)$$

where $U_{\bar{r}}(t)=\sum_{k=1}^{K}\ln\left(D_k-q_k/r_k\right)$, $\bar{r}(t)=\left(r_1(t),\mathrm{L},r_K(t)\right)$, and $\nabla U_{\bar{r}}$ is gradient of $U_{\bar{r}}$. The gradient can be written as

$$\frac{\partial U_r}{\partial r_k}\Big|_{\bar{r}_k}=\frac{q_k}{D_k\bar{r}_k^{-2}-q_k\bar{r}_k} \tag{9}$$

Because the ratio among $\bar{r}(t)$ is equal to the ratio of $\nabla U_{\bar{r}}(t-1)$, the objective function determines the maximum value.

3.3 QoE-Oriented Scheme Based on Dynamic Latency Boundary

Due to the problems mentioned in section 1, we propose a scheme transforming the QoE problem to a QoS problem by setting the latency boundaries for different MOS levels. The scheme computes the latency requirements corresponding to each integral number of MOS, that is, MOS is 1, 2, 3 or 4, for each user, and then selects an achievable MOS level for the latency boundary according to the state of the channel and the services. Finally, the scheme selects N_t users with the largest values of expression (9) and allocates the power iteratively to let the rate ratio of the users approach the gradient. Because this algorithm has no QR-decomposition for selecting users, the time complexity is unrelated to the number of users. According to the expression (4) (5) and (9), the objective function can be rewritten as:

$$\alpha\cdot\frac{q_k}{\left(U_k^{-1}\left(MOS_k\right)-Past_k\right)\bar{r}_k^{-2}-q_k\bar{r}_k} \tag{10}$$

Where $U_k^{-1}()$ maps the MOS to the latency requirement, MOS_k is the achievable MOS level for user *k*. For fairness, let $\alpha=\ln(5/MOS_k)$; therefore, a user with a lower achievable QoE level obtains a higher priority. The steps of the scheme are followed as shown:

01) Initialize $S=\langle\ \rangle$, $H_{K\times N_t}$, *pieces=100,*
02) For k=1,...,K
03) $AchievalRate(k)=f\left(band,SNR,H(k)\right)$
(f() returns the achievable rate if user k get all power)
04) END FOR
05) FOR k=1,...,K
06) MOS=5
07) REPEAT
08) MOS=MOS-1
09) $D_k=DelayBound_k\left(MOS_k\right)-ConsumedTime(k)$
(DelayBound() give the delay requirement to MOS_k*)*
10) UNTIL $D\cdot AchievableRate(k)-Buffering(k)>0\,||\,MOS<1$
11) IF MOS<1
12) gains(k)=0
13) ELSE
14) $gains(k)=\dfrac{\ln(5/MOS_k)\cdot Buffering(k)}{D_k{}^2-Buffering(k)\cdot AchievableRate(k)}$
15) ENDIF
16) ENDFOR
17) Put largest gains (k) into S
18) FOR i=1,...,|S|
19) $Ratio(i)=gains\left(S(i)\right)\Big/\sum_{k\in S}gains(k)$
20) RateRatio(i)=0
21) AssignPowers(i)=0
22) ENDFOR
23) FOR p=1,...,pieces
24) $j=\arg\max_{j\in\{1\mathrm{L}\,|S|\}}\left(Ratio(j)-RateRatio(j)\right)$
25) AssignPowers(j)=AssignPowers(j)+1
26) Update(RateRatio)
27) ENDFOR
28) FOR all $k\in S$ *:*
29) $powers(k)=P\cdot AssignPower(k)/piece$
30) ENDFOR
31) OUTPUT S, powers

Because this scheme does not make QR-decomposition while selecting a user, it only needs $\left(N_t^2-N_t\right)/2$ projection calculations to allocate power, and the time complexity of the algorithm is $\mathrm{O}\left(N_t^2\right)$. If the number of transmit antennas is treated as a constant, the complexity is $\mathrm{O}(1)$.

4. Simulation Result

In this section, we simulated a non-real-time service traffic scenario and an urban wireless scenario to evaluate the performances of the ZF-DPC, QoS and QoE schemes. Many other simulations were performed using real-time traffic models or full buffer models characterized by a constant number of users and unlimited amount of data flow for each user [21]. In our simulation, we adopt a more practical traffic model in which the length of the data flow is finite and the number of users is also constant, which means that each user will start new a service after the reception of the current service completes. Moreover, we suppose the transmit rate of the core network is much faster than the air interface transmission, that is, the BS does not have to wait the data from remote servers. Therefore, we do not need to consider the data arriving rates in the BS.

In our simulation, we consider a BS with four transmitting antennas servicing multiple users with one antenna. The distances between the BS and users are distributed between 30 m and 500 m. The other parameters are listed in Table 1.

Table 1. Simulation Parameters.

Parameter	Value
cell radius	500 m
system band	40 MHZ
noise density	-174 dBm/HZ
speed of users	1 m/s
duration of slot	10 ms
total power	40 w
path loss	34.53+38log10(*D*)
size of video	10~20 Mbits
size of web page	0.5~2.5 Mbits

The types of services include web and short video. The short video starts to play after the reception finishes. The web services are separated into small web page and large web page according to the maximum latency users expect. The number of users ranges from ten to seventy. After one service is completed, the user starts a new service. The simulation lasts one hundred seconds, that is, ten thousand

time slots. According to the type of service, we map latency to the user experience.

In this simulation, we adopt the QoE utility function recommended by [18] to measure the web service:

$$MOS = \frac{4 \cdot \left(\ln(D) - \ln(0.005 \cdot Max + 0.24)\right)}{\ln\left((0.005 \cdot Max + 0.24)/Max\right)} + 5 \tag{11}$$

Where *MAX* is the maximum latency users expect, which is set to 30 s and 60 s for small and large web pages, and *D* is the actual latency. The utility function is proposed by T. Hossfeld in [22]:

$$-1.577 \cdot \ln(D + 0.742) + 5 \tag{12}$$

In our simulation, we examine the average MOS of the schemes. The value can be given by

$$\sum_{i=1}^{K} MOS_i / K \tag{13}$$

Where K is the total number of users, MOS_i is the average MOS of user i, and the MOS of each user is the average MOS of all services triggered by this user, which is equal to:

$$MOS_i = \frac{1}{|U_i|} \cdot \sum_{j \in U_i} MOS_j \tag{14}$$

Where U_i is the set of services completed by user i, MOS_j is the Mean Opinion Score of service j in this set.

Another index we adopted is Jain's Fairness Index (FI) [23]. The FI is given by:

$$FI = \left(\sum_{k=1}^{K} Index_k\right)^2 \Big/ \left(K \cdot \sum_{k=1}^{K} Index_k^2\right) \tag{15}$$

Where *Index* is the MOS of user k or the number of services completed by user k.

Fig. 1 plots the values calculated using expression (13) of the three schemes with the number of users ranging from 5 to 70. Because of the high time complexity of ZF-DPC, we only give the results of ZF-DPC using 5 to 40 users. Fig. 1 shows that the average MOS of the scheme with dynamic latency boundary only experiences a slow decline as the number of users increases; its performance is higher than that of the original QoS provisioning scheme. The performance of ZF-DPC, which has the greatest computational complexity, drops sharply.

Jain's fairness index for the average MOS of the users is illustrated in Fig. 2. The results show that the fairness of our proposed scheme experiences no decline as the number of users increases, while the performances of the other two schemes drop significantly. Fig. 3 shows Jain's fairness index for the services completed by each user. the proposed QoE scheme also achieves better fairness, while the original QoS scheme has the worst performance in the three schemes.

The simulation results show that the proposed QoE-aware scheme achieves better QoE fairness because it can adjust the latency boundary according to the user experience and the changing environment. The original QoS scheme cannot adjust the latency boundary during non-real-time service; thus, it cannot adapt and cannot give different priorities to users with different MOS levels. The defects in ZF-DPC are its high computational complexity and lack of QoE recognition.

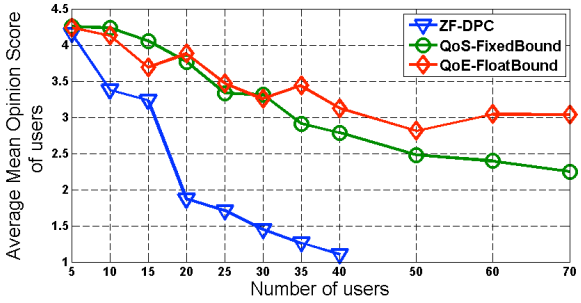

Fig. 1 Average value of users' average MOS

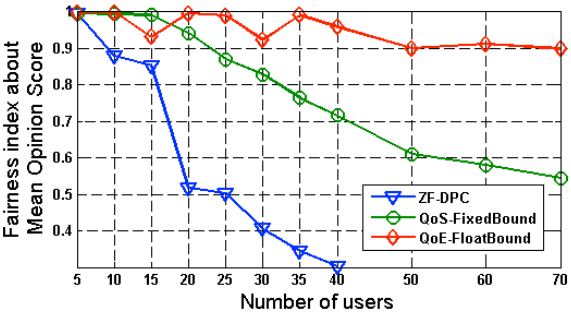

Fig. 2 Fairness about average MOS

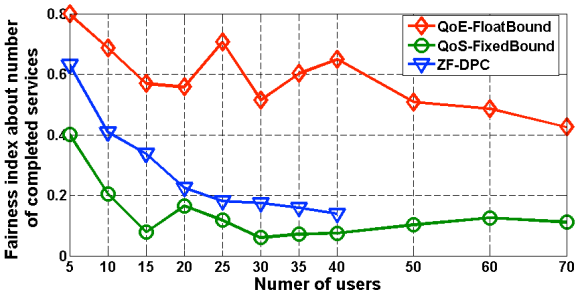

Fig. 3 Fairness about number of completed services

5. Conclusions

We examined three transmission strategies for non-real-time data services in MIMO systems considering QoE fairness. The results showed that the QoE scheduling scheme based on dynamic latency boundary can achieve better average MOS and fairness because our proposed scheme can adjust the latency boundaries according to the state of the channel and the services. The time complexity of this scheme is much lower than traditional schemes because it does not consider channel correlation among users in user selection phase. In this paper, we investigated the schemes with perfect channel state information at the transmitter (CSIT). The case of imperfect CSIT will be an extension of this work.

Acknowledgments

The research reported in this paper was supported in part by a grant from the National High Technology Research and Development Program of China 2012AA111902, and the Fundamental Research Funds for the Central Universities 0800219254.

References

[1]. *H.J. Kushner, "Extensions of proportional-fair sharing algorithms for multi-access control of mobile communications: constraints and bursty data processes", IEEE International Conference on Communications, Seoul, Korea, 2005, pp.3149-3155.*

[2]. *C. Rosa and K.I. Pedersen, "Performance aspects of LTE uplink with variable load and bursty data traffic", IEEE 21st International Symposium on Personal Indoor and Mobile Radio Communications, Istanbul, Turkey, 2010, pp.1871-1875.*

[3]. *Hao Zhu and Guohong Cao, "On improving service differentiation under bursty data traffic in wireless networks", IEEE INFOCOM 2004, Hong Kong, China, 2004, pp.871-881.*

[4]. *J.-B. Landre, A. Saadani, and F. Ortolan, "Realistic performance of hsdpa mimo in macro-cell environment", IEEE 20th International Symposium on Personal, Indoor and Mobile Radio Communications, Tokyo, Japan, 2009, pp. 365–369.*

[5]. *J.-B. Landre, Z. El Rawas, R. Visoz, and S. Bouguermouh, "Realistic performance of lte: In a macro-cell environment", IEEE 75th Vehicular Technology Conference (VTC Spring), Yokohama, Japan, 2012, pp. 1–5.*

[6]. *S. Parkvall, E. Dahlman, A. Furuskar, Y. Jading, M. Olsson, S. Wanstedt, and K. Zangi, "Lte-advanced - evolving lte towards imt-advanced", IEEE 68th Vehicular Technology Conference (VTC Fall), 2008, pp. 1–5.*

[7]. *A. Farajidana, W. Chen, A. Damnjanovic, T. Yoo, D. Malladi, and C. Lott, "3gpp lte downlink system performance", IEEE Global Telecommunications Conference (GLOBECOM), Hawaii, USA, 2009, pp. 1–6.*

[8]. *G. Caire and S. Shamai, "On the achievable throughput of a multiantenna gaussian broadcast channel", IEEE Transactions on Information Theory 49 (7), 2003, pp. 1691–1706.*

[9]. *P. Viswanath and D. Tse, "Sum capacity of the vector gaussian broadcast channel and uplink-downlink duality", IEEE Transactions on Information Theory, 49 (8), 2003, pp. 1912–1921.*

[10]. *S. Vishwanath, N. Jindal, and A. Goldsmith, "Duality, achievable rates, and sum-rate capacity of gaussian mimo broadcast channels", IEEE Transactions on Information Theory, 49 (10), 2003, pp. 2658–2668*

[11]. *H. Weingarten, Y. Steinberg, and S. Shamai, "The capacity region of the gaussian multiple-input multiple-output broadcast channel", IEEE Transactions on Information Theory, 52 (9), 2006, pp. 3936–3964.*

[12]. *T. Yoo and A. Goldsmith, "On the optimality of multiantenna broadcast scheduling using zero-forcing beamforming", IEEE Journal on Selected Areas in Communications, 24 (3), 2006, pp. 528–541.*

[13]. *D. Wu and R. Negi, "Effective capacity: a wireless link model for support of quality of service", IEEE Transactions on Wireless Communications, 2 (4), 2003, pp. 630–643.*

[14]. *J. Tang and X. Zhang, "Quality-of-service driven power and rate adaptation over wireless links", IEEE Transactions on wireless Communications, 6 (8), 2007, pp. 3058–3068.*

[15]. *J. Tang and X. Zhang, "Quality-of-service driven power and rate adaptation for multichannel communications over wireless links", IEEE Transactions on Wireless Communications, 2007, 6 (12), pp. 4349–4360.*

[16]. *L. Zhong and Y. Ji, "Game theoretic qos modeling for joint resource allocation in multi-user mimo cellular networks", IEEE Conference on Wireless Communications and Networking (WCNC), Paris, France, 2012, pp. 1311–1315.*

[17]. *ITU, "Methods for subjective determination of transmission quality", ITU-T Recommendation, 1996.*

[18]. *ITU, "Estimating end-to-end performance in IP networks for data applications", ITU-T Recommendation, 2004.*

[19]. *G. Song and Y. Li, "Utility-based resource allocation and scheduling in ofdm-based wireless broadband networks", IEEE Communications Magazine, 43 (12), 2005, pp. 127–134.*

[20]. *L. Xingmin, T. Hui, S. Qiaoyun, and L. Lihua, "Utility based scheduling for downlink ofdma/sdma systems with multimedia traffic", IEEE Wireless Communications and Networking Conference (WCNC), 2010, pp. 1–6.*

[21]. *P. Ameigeiras, Y. Wang, J. Navarro-Ortis, P. Mogensen, and J. Lopez-Soler, "Traffic models impact on ofdma scheduling design", EURASIP Journal on Wireless Communications and Networking, 2012 (1), 2012, pp. 1-13.*

[22]. *T. Hossfeld, S. Egger, R. Schatz, M. Fiedler, K. Masuch, and C. Lorentzen, "Initial delay vs. interruptions: Between the devil and the deep blue sea", Fourth International Workshop on Quality of Multimedia Experience (QoMEX), Melbourne, Australia , 2012, pp. 1–6.*

[23]. *R. Jain, D. Chiu, and W. Hawe, "A quantitative measure of fairness and discrimination for recourse allocation in shared computer systems", Research Report TR-301 of Digital Equipment Corporation, Maynard, USA, 1984.*